Linear Regression Model for Systolic Blood Pressure

Group 15: Feifei Fu 1006740216, Tianyu Zhang 1006740421, Wenqing Liang 1006739709

2022-12-2

Job Destribution and Description:

Name	Student Number	Job Description
Feifei Fu	1006740216	Cover page, 1/3 of the model analysis
Tianyu Zhang	1006740421	Background and Significance, 1/3 of the model analysis
Wenqing Liang	1006739709	Exploratory data analysis, 1/3 of the model analysis

Introduction

Background: Blood pressure is measured by two values. One is called systolic blood pressure, measuring the pressure in your arteries when your heart beats and the other one is Diastolic blood pressure. (Cologne, 2022)

Goal: Determine which are factors have impact on systolic blood pressure (SBP)

Core Analyses:
1. Correlation matrix
2. Step-wise Regression based on AIC
3. Cross validation
4. Test for outlying : studentized deleted residual, leverage test
5. Test for influence points: DFFITS, cook's distance, DFBETAS

How your data was cleaned (include the reasons why we choose some of the variables):
By doing the research, we found that Jun Miyata and other co-writers stated that the systolic blood pressure which measures the pressure in the arteries when the heart beats is affected more by objective factors, unlike blood pressure which can be affected by many subjective factors in the essay 'Association between high systolic blood pressure and objective hearing impairment among Japanese adults: a facility-based retrospective cohort study'. (Miyata et al., 2022)

Thus, we should try to remove the variables that are more affected by subjective factors or are easily to be unstable, such as income and stress level.

Therefore we should keep the following variables after cleaning the data:

(sbp, ismale, smoke_habit, exercise, age, weight, height, alcohol, trt, bmi, stress, salt, overwt).

Description of Dataset

Cleaning the Blood Pressure data and Include descriptive statistics for each one of the variables

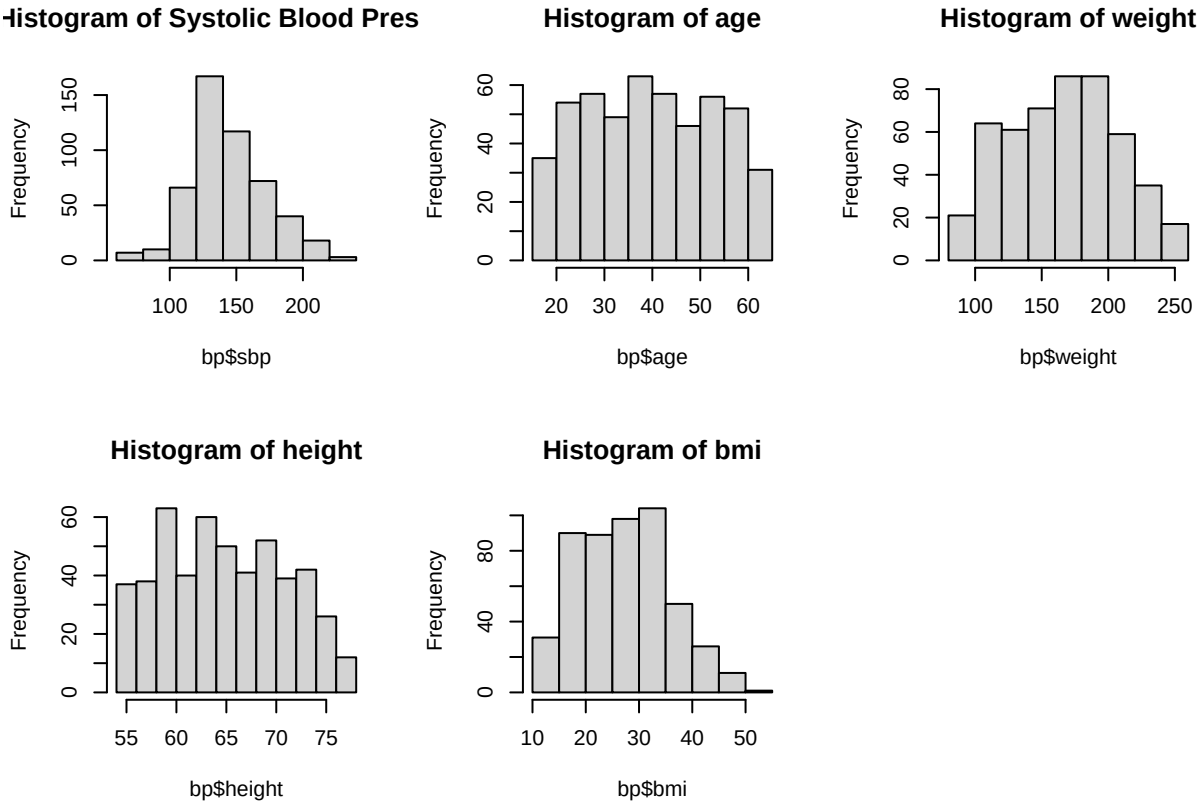
And check the distribution of the response variable

Showing all critical value of y and all covariables such as mean, median and sd:

Variables	Descriptions	Min	Q1	Median	Mean	3rd Q3	Max	SD
sbp	Systolic Blood Pressure (SBP)(Continuous)	67.0	130.0	140.5	145.0	162.2	224.0	28
age	Continuous variable (age years)	18.0	28.0	40.0	52.0	52.0	64.0	13
weight	Continuous variable (lbs).	90.0	133.0	168.0	166.6	198.0	249.0	41
height	Continuous variable (inches)	54.0	60.0	65.0	65.33	70.0	77.0	6
bmi.	Continuous variable: (Weight/Height^2) x 703	11.0	21.0	27.0	27.66	33.0	53.0	9

Showing all the histogram of response variable and covariables:

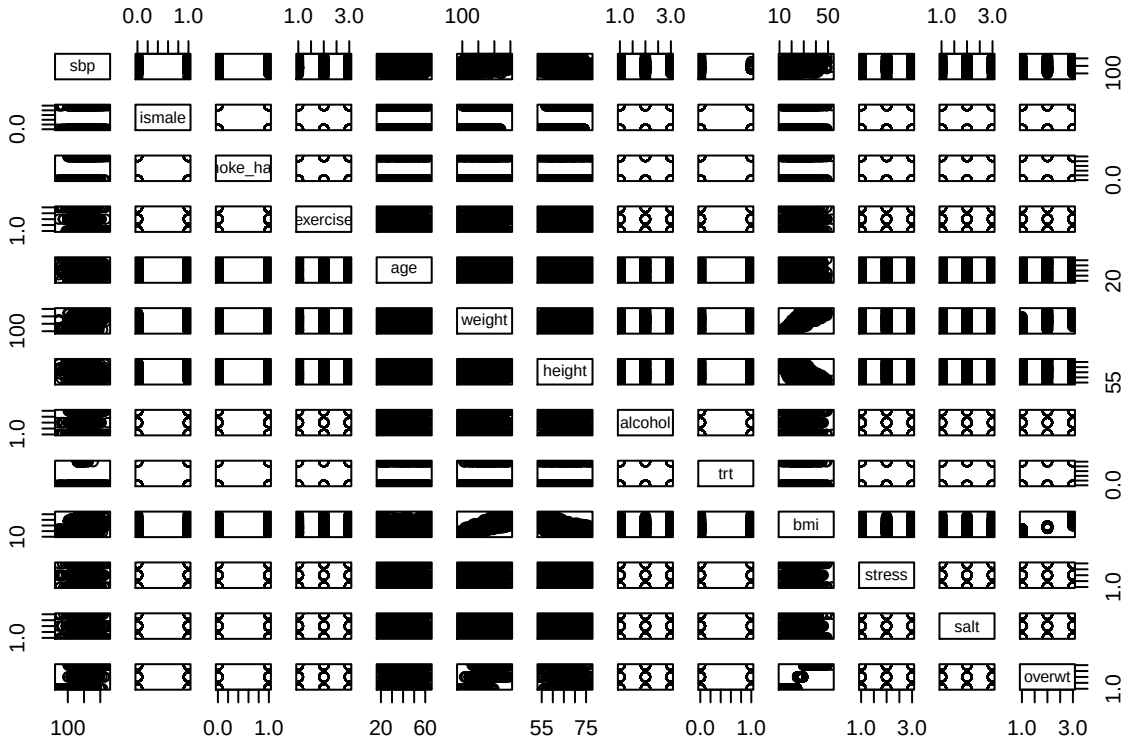
We have dummy variable such as muti-levels of gender,smoke,exercise... We ignore them first:



And for the categorical variables: They are all normal with no obvious missing data or error when we cleaning the data in the first part of the report, we choose not to show the Histograms.

Check the distribution of the response variable and explore the relationships between response and explanatory variables (also between explanatory variables themselves):

Create the scatter plot matrix:



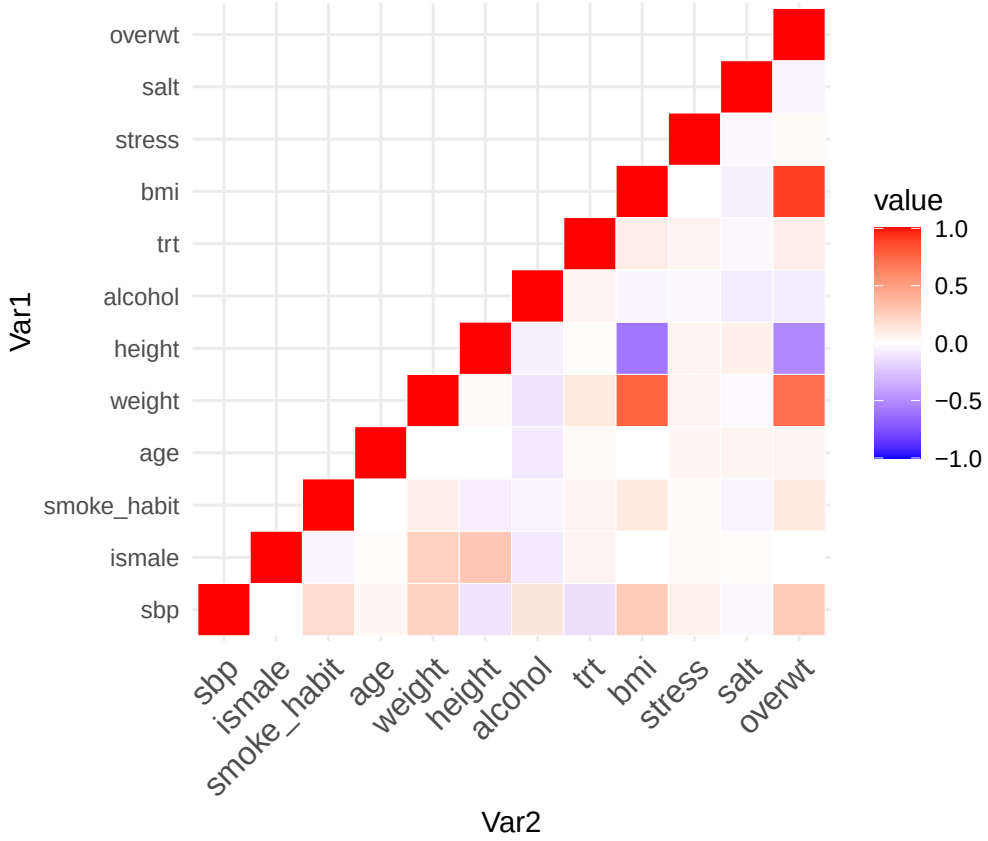
We can also see this by heat map of ggplot:

```
##
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':
##
##     smiths
```

	sbp	ismale	smoke_habit	age	weight	height	alcohol	trt	bmi
sbp	1	0	0.19	0.04	0.23	-0.12	0.13	-0.13	0.27
ismale	NA	1	-0.04	0.01	0.24	0.29	-0.09	0.06	0.00
smoke_habit	NA	NA	1.00	0.00	0.08	-0.07	-0.05	0.06	0.11
age	NA	NA	NA	1.00	0.00	0.00	-0.09	0.03	0.00
weight	NA	NA	NA	NA	1.00	0.03	-0.11	0.12	0.77
height	NA	NA	NA	NA	NA	1.00	-0.06	0.01	-0.59
alcohol	NA	NA	NA	NA	NA	NA	1.00	0.06	-0.04
trt	NA	NA	NA	NA	NA	NA	NA	1.00	0.09
bmi	NA	NA	NA	NA	NA	NA	NA	NA	1.00
stress	NA	NA	NA	NA	NA	NA	NA	NA	NA
salt	NA	NA	NA	NA	NA	NA	NA	NA	NA

```
## overwt       NA     NA          NA   NA     NA     NA      NA    NA    NA
##              stress  salt overwt
## sbp           0.07 -0.03   0.27
## ismale        0.03  0.01   0.00
## smoke_habit   0.03 -0.05   0.12
## age           0.04  0.05   0.05
## weight        0.06 -0.02   0.72
## height        0.06  0.08  -0.51
## alcohol      -0.03 -0.08  -0.08
## trt           0.06 -0.03   0.09
## bmi           0.00 -0.06   0.89
## stress        1.00 -0.03   0.03
## salt            NA  1.00  -0.05
## overwt          NA    NA   1.00
```



Some of the cor is pententially large –> May have Multicollinearity: 1.(weight,overwt)
2.(weight,bmi)
3.(overwt, bmi) And we can find that:
1.cor(sbp, overwt) is quiet small
2.cor(sbp, bmi) is quiet small
3.cor(sbp, weight) is quiet small 4.Weight line showed up two problems, this variable is worse than other variables 5.overwt column showed up two problems, this variable is worse than other variables Therefore we may drop those variables in the following model selection steps, especially weight and overwt.

Explore the relationships between response and explanatory variables:

By the scatter plot matrix and correlation matrix:

Variables	Cov	Relationship
“sbp” and “ismale”	0.002338999	NO linear association.
“sbp” and “smoke_habit”	0.193448533	weak positive linear association
“sbp” and “exercise”	-0.14537399	weak negative linear association
“sbp” and “age”	0.037463336	NO linear association.
“sbp” and “weight”	0.230277555.	weak positive linear association
“sbp” and “height”	-0.116917759	weak negative linear association
“sbp” and “alcohol”	0.13261708	weak positive linear association
“sbp” and “trt”	-0.12577842	weak negative linear association
“sbp” and “bmi”	0.2666692719	weak positive linear association
“sbp” and “stress”	0.06682121	NO linear association
“sbp” and “salt”	-0.02923472	NO linear association
“sbp” and “overweight”	0.266564463	weak positive linear association

And by the 12 observation based on the response and explanatory variables, we can see that the most associated explanatory variable is bmi which has cor(sbp, bmi) = 0.2666692719, and it is positive linear association. And there is no association between sbp and gender,stress,salt,stress.

Explore the relationships between explanatory variables themselves:

And for other relationship between other explanatory variables:
We can just to find the relationship from the heat-map of corrlation matrix:
1. The cor(gender(ismale) and other explanatory variables) are showed with white-similar color except weight and height:
therefore there is no linear association with all other explanatory variables except weight and height.
(height and ismale), (weight and ismale) are weakly associated.
2. The cor(smoking and other explanatory variables) are showed with white-similar color:
therefore there is no linear association with all other explanatory variables.
3. The cor(age and other explanatory variables) are showed with white-similar color:
therefore there is no linear association with all other explanatory variables.
4. The cor(weight and other explanatory variables) are showed with white-similar color except bmi and overwt:
therefore there is no linear association with all other explanatory variables except bmi and overwt.
(weight and bmi), (weight and overwt) are highly associated.
5. The cor(height and other explanatory variables) are showed with white-similar color except bmi and overwt:
therefore there is no linear association with all other explanatory variables except bmi and overwt.
(height and bmi), (height and overwt) are moderately associated.
6. The cor(alcohol and other explanatory variables) are showed with white-similar color:
therefore there is no linear association with all other explanatory variables.
7. The cor(trt and other explanatory variables) are showed with white-similar color:
therefore there is no linear association with all other explanatory variables.
8. The cor(bmi and other explanatory variables) are showed with white-similar color except overwt:
(bmi and overwt) are highly associated.
9. The cor(stress and other explanatory variables) are showed with white-similar color:
therefore there is no linear association with all other explanatory variables.
10. The cor(salt and other explanatory variables) are showed with white-similar color:
therefore there is no linear association with all other explanatory variables.

Building Model and Model Validation

Main effect model:

We choose all variables to build the model fit1. Then, we apply stepwise regression based on AIC to fit1 to build model_1. After calculating the value of p, R^2_{adj}, C_p, AIC, BIC, $PRESS$, we create Table1.

Table1: Comparing fit1 and model_1 based on the value of p, R^2_{adj}, C_p, AIC, BIC, $PRESS$

	# of predictors	Adjusted R^2	C_p	AIC	BIC	$PRESS$
fit1	17	0.1821	18	4670.134	4750.212	331426.4
Model_1	11	0.1875	8.757558	4660.987	4715.776	325466.7

We can see that
1.Adjusted R Squared increases
2.AIC, BIC and PRESS decreases So model_1 is better

Then we compare model_1 with another model with limited variables, this part comes from the based part about why we choose those variables. We name the reduced model as fit2. Then, we apply stepwise regression based on AIC to fit2 to build model_2. After calculating the value of p, R^2_{adj}, C_p, AIC, BIC, $PRESS$, $|C_p - p'|$, we create Table2.

Table2: Comparing fit2, model_1 and model_2 based on the value of p, R^2_{adj}, C_p, AIC, BIC, $PRESS$, $|C_p - p'|$.

	# of predictors	Adjusted R^2	C_p	AIC	BIC	$PRESS$	
fit2	12	0.1777	15.6195	4667.976	4726.981	329923.3	2.6195
Model_1	11	0.1875	8.757558	4660.987	4715.776	325466.7	3.242442
Model_2	9	0.1818	10.16681	4662.535	4708.896	326409.6	0.16681

By comparing the four models, we found that fit2 has the biggest AIC, BIC, PRESS value, so we do not consider fit2. Since the values of AIC are similar and have small difference, we will compare other values. Then we can see that the model_2 has the smallest BIC and it's Cp is the nearest to P'. Therefore we think model_2 is better than fit2 and model_1.

Then we add interaction terms:
1.between bmi and smoke_habit.
2.between bmi and trt 3.between bmi and exercise

Interaction Model with stepwise selection:

We add interaction terms bmi:smoke_habit + bmi:factor(trt) + bmi:factor(exercise) to model_2 and apply stepwise selection to it, We name the new model as model3 after calculating the value of p, R^2_{adj}, C_p, AIC, BIC, $PRESS$, we create Table3.

Table3: Comparing model_2 and model_3 based on the value of p, R^2_{adj}, C_p, AIC, BIC, $PRESS$.

	# of predictors	Adjusted R^2	C_p	AIC	BIC	$PRESS$
Model_2	9	0.1818	10.16681	4662.535	4708.896	326409.6
Model_3	11	0.2092	-10.17362	4647.468	4702.258	315880.6

We can see that

1.Adjusted R Squared increases
2.AIC, BIC and PRESS decreases

So model_3 is better

Added variable Plot:
Since all variable have no missing datas, and there is no curvilinear band in those plots, so we don't consider power model

Then we can find the summary estimators of the final model and indicate the final model:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	82.0159012	18.6476827	4.3981819	0.0000134
bmi	1.0430346	0.2324657	4.4868324	0.0000090
factor(smoke_habit)1	11.6059835	2.2608872	5.1333758	0.0000004
factor(trt)1	17.3666505	10.3071373	1.6849150	0.0926443
factor(alcohol)2	1.0662138	2.7779419	0.3838143	0.7012832
factor(alcohol)3	12.0084383	2.7716516	4.3325930	0.0000179
factor(exercise)2	-14.7124201	9.4293411	-1.5602808	0.1193418
factor(exercise)3	-32.3534832	9.0328311	-3.5817656	0.0003755
height	0.4968010	0.2283287	2.1758147	0.0300482
bmi:factor(trt)1	-1.0706256	0.3417668	-3.1326197	0.0018367
bmi:factor(exercise)2	0.1722332	0.3166029	0.5440038	0.5866873
bmi:factor(exercise)3	0.8062522	0.3188938	2.5282777	0.0117766

The final regression eqution:
lm(formula = sbp ~ bmi + factor(smoke_habit) + factor(trt) + factor(alcohol) + factor(exercise) + height + bmi:factor(trt) + bmi:factor(exercise), data = bp_original)

Interpretations:
$\hat{\beta_0} = 82.015901$ represents the average SBP for a person with weight = 0, height = 0 and also this person is a non-smoker and not get trt, alcohol use is low, Exercise level is low is 82.015901 $\hat{\beta_1} = 1.043034$ represents keeping all other variables constant, as BMI increase by 1, average SBP increase by 1.043034
$\hat{\beta_2} = 11.6059835$ represents keeping all other variables constant, SBP of smoker is expected greater than the SBP of non-smoker by 11.6059835
$\hat{\beta_3} = 17.366650$ represents keeping all other variables constant, SBP of the person got trt is expected greater than the SBP of the person did not get trt by 17.366650
$\hat{\beta_4} = 1.066213$ represents keeping all other variables constant, SBP of alcohol use is Medium level is expected greater than the SBP of alcohol use is low level by 1.066213
$\hat{\beta_5} = 12.008438$ represents keeping all other variables constant, SBP of alcohol use is high level is expected greater than the SBP of alcohol use is low level by 12.008438
$\hat{\beta_6} = -14.712420$ represents keeping all other variables constant, SBP of Exercise level is medium is expected lower than the SBP of Exercise level is low by 14.712420
$\hat{\beta_7} = -32.353483$ represents keeping all other variables constant, SBP of Exercise level is high is expected lower than the SBP of Exercise level is low by 32.353483
$\hat{\beta_8} = 0.496801$ represents keeping all other variables constant, as height increase by 1, average SBP increase by 0.496801
$\hat{\beta_9} = -1.070625$ represents keeping all other variables constant, the difference between the slope of bmi (same or fixed) for "the person got trt" and the slope of bmi (same or fixed) for "the person did not get trt". And that difference in slope is -1.070625
$\hat{\beta_1 0} = 0.1722332$ represents keeping all other variables constant, the difference between the slope of bmi (same or fixed) for Exercise level is medium and the slope of bmi (same or fixed) for Exercise level is low. And that difference in slope is 0.1722332

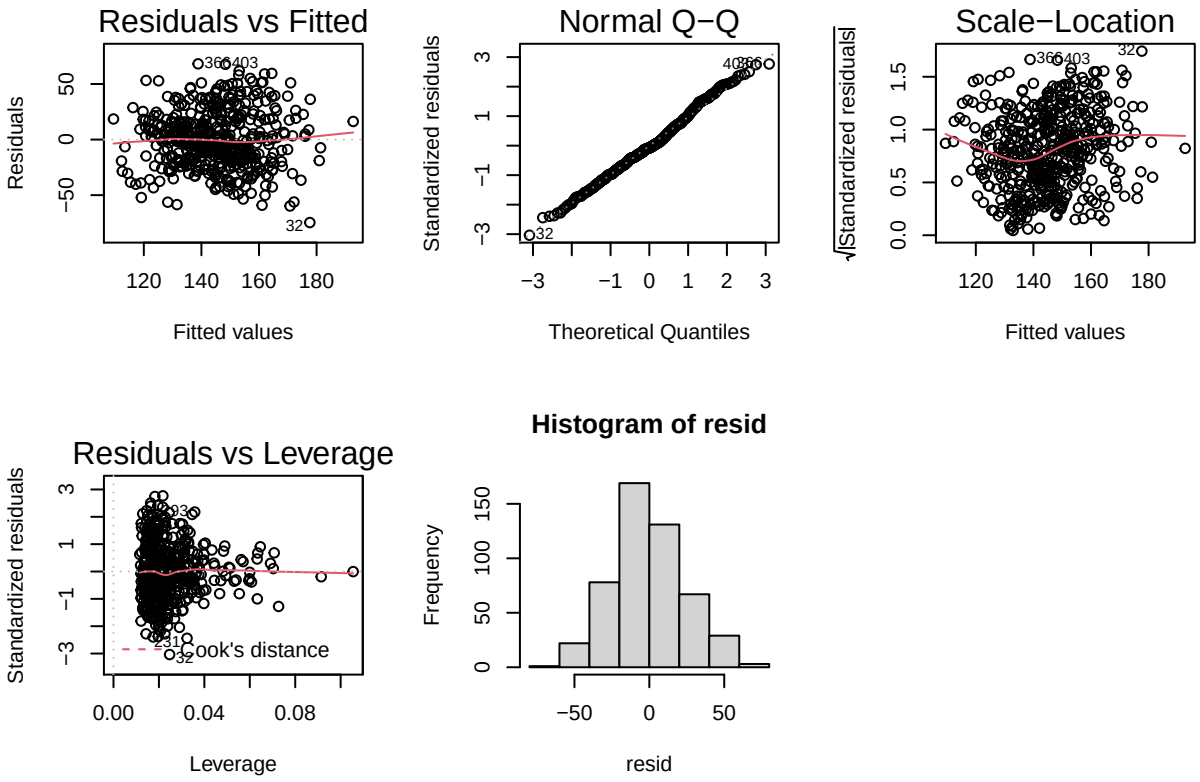
$\hat{\beta_1}0 = 0.806252$ represents keeping all other variables constant, the difference between the slope of bmi (same or fixed) for Exercise level is high and the slope of bmi (same or fixed) for Exercise level is low. And that difference in slope is 0.806252

Cross-validation

Finding: 1.MSPR = 228194.4 and the MSE = 594.7
2.MSPR is far away from MSE, so the validation told us that the model is still important and useful, but the "out of sample" predictive power of the final model is weak.

Model Diagnostics

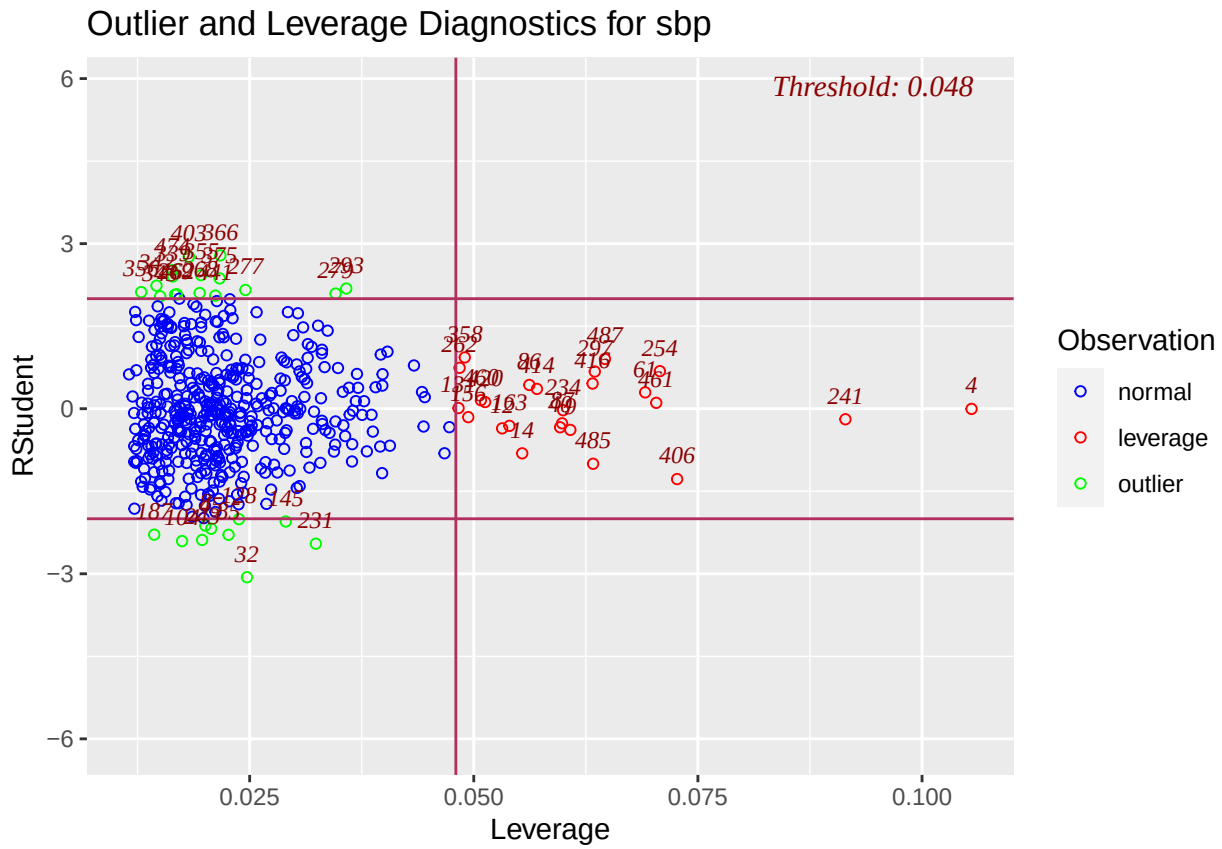
Line assumption



There is no pattern in the residual plot, the QQ line of residual plot is close to the line(y=x) and the histogram of residual plot looks normal. So, the assumptions of linearity, normality, independent and equal variance are satisfied. Therefore, we don't need to do the Box-Cox transformation.

Outerliers and influential observations

Check the outerliers by graph:



By the R diagnostic graphs, the outliers and leverage is shown above.

Check the influential observations by hand:

The influential observations, DFFITS perspective:

```
##  32  45  85 104 128 145 231 243 277 279 293 339 355 366 375 403 406 474
##  32  45  85 104 128 145 231 243 277 279 293 339 355 366 375 403 406 474
```

The influential observations, cook distance:

```
## named integer(0)
```

The influential observations, DFBETAS perspective:

```
##   [1]   32   45   66   85  115  158  183  235  263  269  275  292  302  310  379
##  [16]  399  403  471  502  601  603  604  628  648  657  666  692  725  762  779
##  [31]  790  839  855  867  875  884  887  908  927  962 1008 1045 1145 1187 1188
##  [46] 1243 1247 1268 1329 1339 1355 1356 1366 1403 1446 1486 1514 1601 1915 1987
##  [61] 1996 2052 2094 2107 2187 2207 2227 2243 2247 2279 2293 2309 2329 2368 2441
##  [76] 2446 2469 2552 2594 2607 2625 2687 2707 2727 2743 2747 2768 2787 2798 2842
##  [91] 2843 2866 2869 2882 2958 2962 2964 2974 3101 3126 3128 3157 3166 3225 3231
## [106] 3269 3274 3279 3303 3313 3318 3348 3375 3406 3412 3466 3576 3585 3601 3628
## [121] 3645 3657 3666 3725 3740 3777 3779 3793 3841 3866 3869 3875 3893 3927 3942
## [136] 3965 3985 3987 4011 4029 4052 4099 4104 4112 4147 4225 4245 4247 4268 4334
## [151] 4342 4355 4378 4446 4458 4482 4532 4731 4754 4906 5032 5045 5053 5068 5149
## [166] 5165 5243 5329 5358 5381 5382 5391 5404 5441 5446 5469 5486 5508 5532 5545
## [181] 5607 5625 5635 5714 5716 5797 5798 5822 5829 5861 5881 5882 5946 5974 5986
```

- From the DFFITS perspective, there are 18 influential observations, which are item: 32 45 85 104 128 145 231 243 277 279 293 339 355 366 375 403 406 474
- From the cook distance perspective, there are 0 influential observations.
- From the DFBETAS perspective, there are 195 influential observations.

Check the leverages by hand:

There are 19 leverage points, which are item: 4 9 12 14 40 61 86 87 163 234 241 254 297 406 414 416 461 485 487

Check the outlying points by hand:

There is 0 outlying points in Y observations.

Conclusion

Summarize findings:

Those factors are gonna make true and most effective impact on systolic blood pressure:
1. bmi
2. Smoking Status
3. Treatment (for hypertension)
4. Alcohol use
5. Exercise level
6. Height
7. Interaction term between bmi and treatment
8. Interaction term between bmi and exercise

limitations of our study:

1. validation step told us the“out of sample” predictive power of the final model is weak
2. Too much influential observation (outliers and influential point)
3. Have the multicollinearity problem result from interaction term

Suggest future directions/extensions:

1. Try to collect more specific data, increase the sample size, we can also use Time series such as SARIMA model to predicts future values
2. Further investigate influence observations, try to deal with them and fit a better model

Reference

1.Miyata, J., Umesawa, M., Yoshioka, T. et al. Association between high systolic blood pressure and objective hearing impairment among Japanese adults: a facility-based retrospective cohort study. Hypertens Res 45, 155–161 (2022). https://doi.org/10.1038/s41440-021-00737-8/

2.Kristine Thorndyke (2022). Understand Subjective vs Objective Data: https://testprepnerds.com/nclex/subjective-vs-objective-data/

3.Cologne (2022).Institute for Quality and Efficiency in Health Care (IQWiG): https://www.ncbi.nlm.nih.gov/books/NBK279251/